UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

Relatórios Técnicos
do Departamento de Informática Aplicada
da UNIRIO
n° 0002/2019

Extração de Informações Baseada em Ontologia: Oportunidades e Desafios

Jônatas Castro dos Santos
Sean Wolfgand Matsui Siqueira

Departamento de Informática Aplicada

UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
Av. Pasteur, 458, Urca - CEP 22290-240
RIO DE JANEIRO – BRASIL

Extração de Informações Baseada em Ontologia: Oportunidades e Desafios*

Jônatas Castro dos Santos Sean Wolfgand Matsui Siqueira
Depto de Informática Aplicada – Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
{jonatas.santos, sean}@uniriotec.br

Abstract. Ontology-based Information extraction systems (OBIE) allow the automatic ex-traction of concepts in natural language texts at various domains of society. Several OBIE's approaches have arisen by proposing many alternatives for ex-traction methods. Therefore, this work presents a systematic mapping on the recent OBIE approaches. We propose a generic process that considers the population of ontology and we also identify opportunities and challenges for OBIE, including questions about language and translation, OBIE architecture performance and ontology quality.

Keywords: ontology-based information extraction; natural language processing

Resumo. Os sistemas de extração de informação baseada em ontologia (OBIE) permitem a extração automática de conceitos em textos de linguagem natural em diversos domínios da sociedade. Muitas Abordagens de OBIE têm surgido propondo múltiplas variantes para os métodos de extração. Por isso, este trabalho apresenta um mapeamento sistemático sobre as abordagens de OBIE recentes. Propomos um processo genérico que leva em conta o povoamento de ontologias e identificamos oportunidade e desafios para a área de OBIE, entre eles, questões sobre linguagem e tradução, desempenho da arquitetura OBIE e qualidade da ontologia.

Palavras-chave: extração de informações baseada em ontologia; processamento de linguagem natural

* Trabalho patrocinado pela CNPq: 03/2019 - 02/2022

1 Introdução

Extração da informação (*Information Extraction* - IE) é o processo que envolve a recuperação automática de certos tipos de informação a partir de documentos de texto não-estruturados ou semiestruturados (Russell & Norvig, 2009) (Wimalasuriya & Dou, 2010). IE é considerada como parte do domínio de inteligência artificial e se preocupa com a aquisição de conhecimento através da busca de ocorrências de uma classe particular de objetos e suas relações dentro de um domínio.

O processo de IE ajuda muitas áreas que necessitam de mecanismos eficazes para automatizar a captura de informações de textos em linguagem natural, relatórios e bases não-estruturadas. Sem mecanismos automáticos, essas áreas estão fadadas a esforços manuais para extração de informações. São exemplos de domínios que buscam IE: médico (De Silva, Dou, & Huang, 2017; Viani et al., 2018), engenharia civil (Zhou & El-Gohary, 2017), robótica (Ali et al., 2017), jurídico (de Araujo, Rigo, & Barbosa, 2017), redes de computadores (Martinez, Yannuzzi, de Vergara, Serral-Gracia, & Ramirez, 2015) etc.

Apesar de grande parte das fontes processadas por sistemas de IE ser composta por textos em linguagem natural, não é de inteira responsabilidade de IE a compreensão da linguagem natural (Mannai, Karâa, & Ghezala, 2018; Rau, Jacobs, & Zernik, 1989). Para esse fim, os sistemas de IE utilizam ferramentas de Processamento de Linguagem Natural (NLP). Esta área é um subtópico de Entendimento de Linguagem Natural que inclui, além do entendimento de textos em linguagem natural, a compreensão e síntese da voz humana. Exemplos de aplicações de NLP são os *chat bots* (Shawar & Atwell, 2004) e os assistentes pessoais inteligentes (Comerford, Frank, Gopalakrishnan, Gopinath, & Sedivy, 2001). As técnicas de NLP mais utilizadas em IE são: *part-of-speech (POS) tagging*, análise morfossintática, *tokenização*, lematização etc. (Manning et al., 2014) (Zhou & El-Gohary, 2017).

Também é importante diferenciar IE de recuperação da informação (IR), processo que visa recuperar documentos (ou parte deles) a partir de uma consulta a uma grande coleção de documentos. Um bom exemplo de sistemas de IR são os mecanismos de busca populares, como Google e BING. A principal diferença de IE e IR é que o último efetua análise do texto e provê a informação por conta própria, enquanto o primeiro funciona apenas como um buscador de documentos (De Silva et al., 2017).

Extração de informação baseado em ontologia (Ontology Based Information Extraction - OBIE) é um subtópico de extração de informação onde ontologias são utilizadas pelo processo de IE para diversos fins (Wimalasuriya & Dou, 2010). Ontologia é um conceito genérico que permeia por várias áreas de conhecimento (Antoniou & van Harmelen, 2004). No contexto de Ciências da Computação e Informação, uma ontologia define um conjunto de representações primitivas com objetivo de modelar um domínio de conhecimento ou discurso (Gruber, 2007).

Daya C. Wimalasuriya e Dejing Dou, autores do estudo mais citado na área (Wimalasuriya & Dou, 2010), definiu OBIE como "um sistema que processa dados não-estruturados ou semiestruturados de texto em linguagem natural através de um mecanismo guiado por ontologias para extrair certos tipos de informações e apresentar uma saída utilizando ontologias". Como constatado pelos autores, as ontologias podem guiar o processo de OBIE, mas também podem representar o resultado do processo do OBIE. Esta saída contém um conjunto de entidades, instâncias, relações e regras contidas no dado processado.

Ao longo da década de 2010, novas abordagens de OBIE foram propostas. Pode-se notar a utilização de sistemas de OBIE para uma diversidade de domínios. Porém, poucos estudos se preocupam em compilar essas abordagens para organizar ideias, comparar e identificar semelhanças entre elas e promover a reutilização de ontologias que apoiem o OBIE. Percebe-se ainda que as ontologias podem ser utilizadas em diferentes fases do processo do OBIE. Entretanto, em meio a tantas abordagens, ainda não há um consenso quanto à utilização mais adequada de ontologia para apoiar o processo de OBIE.

Este estudo visa a apresentar e relacionar abordagens de OBIE com o foco de ontologia dentro do processo de OBIE, bem como, identificar oportunidades e desafios em OBIE. Foi realizado um mapeamento sistemático dos estudos mais recentes para selecionar as abordagens relevantes de OBIE. Relacionar essas abordagens em um eixo de comparação pode ajudar futuros pesquisadores no processo de criação e/ou utilização de ontologias para melhor suportar sistemas de IE.

Com intuito de abordar esse tema de modo sucinto, na sessão seguinte, iremos explicar o processo de OBIE. Na seção 3, explicitaremos nossa pesquisa de abordagens atuais de OBIE e discutiremos oportunidades e desafios em 4. Na seção final, apontaremos as conclusões sobre o estudo.

2 O processo de OBIE

Para que um sistema seja considerado de OBIE, deve-se levar em conta alguns fatores-chave. Esses fatores diferenciam os sistemas de OBIE de sistemas de IE comuns (Wimalasuriya & Dou, 2010) (Fudholi, Rahayu, & Pardede, 2016):

- Processa dados não-estruturados ou semiestruturados em linguagem natural;
- Apresenta a saída utilizando ontologias;
- Utiliza um processo de IE guiado por ontologias.

Já há certo consenso sobre esses requisitos, a saber que a maioria das abordagens que se auto denominam OBIE possui essas características. Porém, o que mais diverge entre as abordagens é o modo em que as ontologias são utilizadas. Para entender essas diferenças, evidenciamos um processo genérico em que maior parte dos estudos conduz OBIE (Figura 1).

No subprocesso de **pré-processamento**, são aplicadas diversas técnicas de NLP sobre o texto bruto de entrada (e.g. (Manning et al., 2014)). Esta etapa gera um conjunto de dados semiestruturados com marcações (anotações).

Nesta etapa, várias técnicas de processamento de texto podem ser aplicadas para efetuar limpeza, tratamento, filtro e marcação dos dados. O resultado deste processo, geralmente, são documentos semiestruturados com anotações léxicas, morfológicas, sintáticas e semânticas, na maioria das vezes, baseadas em regras linguísticas. Em (Lima, Espinasse, & Freitas, 2017), documentos XML, que foram gerados a partir de regras linguísticas, são a saída deste subprocesso. Outro formato bastante comum é o Stanford XML (De Silva et al., 2017), fazendo referência ao conjunto de ferramentas Stanford Core NLP proposto em (Manning et al., 2014). Em alguns casos, como em (Zhou & El-Gohary, 2017), ainda há um pré-filtro do texto para evitar esforços desnecessários de processamento.

A partir desses dados marcados, o subprocesso de **extração da informação** compõe o núcleo de procedimentos que efetivamente extrai conhecimento dos dados. Em sistemas

de OBIE, esses procedimentos de IE são guiados por uma ontologia que, geralmente, é detentora do conhecimento do domínio. (Wimalasuriya & Dou, 2010) explicita que nenhum método de IE novo é criado, mas os métodos devem ser orientados para identificar os componentes de uma ontologia.

Existem pelo menos duas classificações de métodos de IE que são muito utilizados em OBIE (Zhou & El-Gohary, 2017): IE baseado em regras e IE baseado em abordagens de aprendizado de máquina (*machine learning - ML*).

Os métodos de IE baseado em regras necessitam de mão-de-obra humana para extrair *features* de uma parte pequena do texto, definir padrões com relação a essas *features* e desenvolver regras de extração com base nesses padrões (Zhou & El-Gohary, 2017). Esses métodos têm grande intercessão com as técnicas explicitadas no subprocesso de pré-processamento.

Já os métodos de IE baseado em ML, necessitam de um grande conjunto de dados anotados (de forma automática ou não) com *features* e informações a serem extraídas. Esses dados são denominados *training* data. Então, um algoritmo de ML é aplicado nesses dados para inferir regras de extração de informações de forma automática. Quando o conjunto de dados é gerado a partir de esforço humano, esses métodos são denominados *supervisionados* (e.g. *Support Vector Machines, Hidden Markov Models, Conditional Random Fields).* Quando a abordagem não necessita de um conjunto de dados para inferir regras, chamamos de métodos de ML *não-supervisionados* (Zhou & El-Gohary, 2017).

O subprocesso de **construção de ontologia** não é necessariamente obrigatório para um OBIE. Geralmente, ocorre antes da implementação do sistema, mas pode ser que a abordagem utilize ou modifique uma ontologia de domínio já existente. Os detalhes sobre construção de ontologia não fazem parte do escopo deste trabalho. Porém, é importante evidenciar que algumas abordagens declaram certos requisitos de uma ontologia (e.g. (Martinez et al., 2015)).

A etapa de **povoamento de ontologia** (*ontology population*) é considerada uma etapa fundamental para garantir enriquecimento semântico (Fudholi et al., 2016). Entretanto, nem todos os estudos se preocupam em prover essa funcionalidade. A ideia é que o subprocesso principal extraia instâncias conhecidas pela ontologia.

(Fudholi et al., 2016) consideram enriquecimento e povoamento de ontologias como o principal objetivo de sistemas OBIE. A ideia desse subprocesso é preencher a ontologia com instâncias dos conceitos extraídos. Estas instâncias podem ser de classes, entidades ou relações da ontologia de domínio. Teoricamente, toda abordagem OBIE é capaz de realizar esta etapa, uma vez que os sistemas extraem instâncias de uma ontologia de domínio.

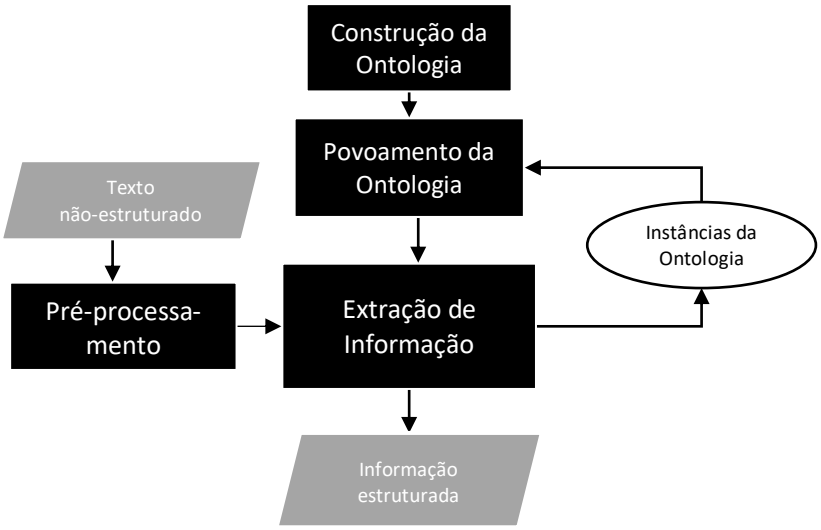

Figura 1 - Proposta de processo genérico de abordagens OBIE

3 Abordagens de OBIE recentes

Um dos objetivos deste trabalho é apresentar um panorama das abordagens mais recentes de OBIE. Para isso, efetuamos um mapeamento sistemático buscando por artigos de periódicos e revistas nas bases Elsevier's Scopus[1], IEEE Xplorer[2] e ACM Digital Library[3].

Em nossa pesquisa, utilizamos a seguinte *string* de busca:

```
("ontology-based" OR "ontology-driven") AND "information extraction"
```

A partir dessa busca, selecionamos estudos de 2016 a 2019, utilizando os seguintes critérios de inclusão:

(1) classificação do QUALIS A1, A2 ou B1;

(2) estudo deve explicitamente apresentar uma abordagem ou sistema de OBIE.

Excluímos do escopo da pesquisa estudos que focassem apenas em sistemas de recuperação da informação, mineração semântica de dados ou processamento de linguagem natural. Limitamos a listagem final em dez estudos por questões de espaço.

Baseado na tabela de classificação de (Wimalasuriya & Dou, 2010), geramos duas tabelas que sumarizam os estudos selecionados em nossa pesquisa (Tabela 1 e

Tabela 2). Em relação a tabela de (Wimalasuriya & Dou, 2010), separamos a classificação de métodos de IE, passando a diferenciar aqueles baseados em regras dos baseados em ML. Além dessa modificação, acrescentamos as colunas: "povoamento de ontologias", "idioma", "ferramentas/frameworks", "dicionários/*theasaurus*/*gazetters*", "validação da abordagem" e "métricas de avaliação". Consideramos essas informações importantes, uma vez que impactam na forma como ontologias são utilizadas em OBIE.

[1] https://www.scopus.com

[2] http://ieeeaccess.ieee.org

[3] https://dl.acm.org

4 Desafios e oportunidades em OBIE

Buscamos nesta seção evidenciar problemas abertos na área de OBIE que foram elencados a partir da pesquisa explicada na seção anterior. Destacamos seis tópicos pertinentes ao contexto de OBIE: métodos de IE utilizados em OBIE (4.1), povoamento de ontologias (4.2), linguagem e tradução (4.3), combinação de ontologias (4.4), desempenho (4.5) e qualidade da ontologia (4.6).

4.1 Métodos de IE utilizados em OBIE

Vale ressaltar que, enquanto os estudos anteriores proviam mais OBIE através métodos baseados em regras, os estudos mais recentes têm utilizado métodos baseados em técnicas de ML (Lima et al., 2017; K. Liu & El-Gohary, 2017). Porém, os métodos de IE baseados em ML que lidam com semântica ainda precisam ser mais explorados para suportar sistemas de OBIE (K. Liu & El-Gohary, 2016).

Há uma certa tendência para que os sistemas OBIEs utilizem mais abordagens de ML (Ali et al., 2017; Lima et al., 2017; K. Liu & El-Gohary, 2017) a fim de aumentar a precisão da extração de informação, porém, o desempenho de abordagens ML ainda é um problema complexo (Zhou & El-Gohary, 2017). Por outro lado, muitas abordagens de OBIE se apoiam apenas em métodos de IE baseado em regras (De Silva et al., 2017; Rizvi et al., 2018; Viani et al., 2018).

Outro ponto sobre métodos de IE é se os componentes do método de extração devem fazer parte da ontologia, conforme discutido em (Wimalasuriya & Dou, 2010). Percebe-se que ainda não há consenso sobre isso, pois vemos tanto abordagens que agregam os esses componentes na ontologia (K. Liu & El-Gohary, 2016), como abordagens que não utilizam ontologia para esse fim (Ali et al., 2017; Martinez et al., 2015).

4.2 Povoamento de Ontologia (*Ontology Population*)

Alguns estudos afirmam que todo OBIE pode ser tratado como um sistema de povoamento de ontologia (Fudholi et al., 2016; Lima et al., 2017; Manine, Alphonse, & Bessières, 2008), porém, podemos constatar que, no panorama atual, nem toda abordagem de OBIE faz menção de povoamento de ontologia.

A possibilidade de popular ontologias de forma automática abre um grande potencial para os OBIEs, pois permite que o sistema lide com a extração de um conhecimento que ainda não foi aprendido anteriormente. Esse problema também é conhecido como *enriquecimento semântico* (Fudholi et al., 2016).

Ainda nesse contexto, duas questões ainda pouco exploradas são: (1) o repovoamento contínuo de ontologias com informações mais recentes e (2) validação automática do conhecimento da ontologia (Fudholi et al., 2016).

4.3 Linguagem e tradução

A maioria das abordagens de OBIE e NLP contém domínios na língua inglesa. Apesar dos avanços em IE, poucos estudos apresentam abordagens em outras línguas (de Araujo et al., 2017; Viani et al., 2018). Uma hipótese para isso seria a falta de ferramentas e recursos. (Viani et al., 2018), por exemplo, apresenta uma abordagem para extração de informações guiadas por ontologia em relatórios médicos na língua italiana. O autor menciona a falta de bases anotadas em italiano para que seja possível efetuar a validação

eficaz do modelo. Em (de Araujo et al., 2017), a fonte de dados é o conjunto de documentos jurídicos em português (Brasil).

Alguns estudos também defendem a necessidade de um suporte para estender a abordagem para outros idiomas (Viani et al., 2018). Na maioria dos casos, a própria ontologia, que contém os componentes extratores da OBIE, favorece a extensibilidade de idiomas. Teoricamente, a simples tradução desses componentes seria o suficiente para permitir a utilização do sistema para um novo idioma, porém é sabido que nem todo conceito pode ser traduzido para outro idioma de forma fidedigna.

Como explicitado anteriormente, há falta de ferramentas em outros idiomas. Essas ferramentas incluem os componentes do subprocesso de pré-processamento que levam em conta um conjunto de regras linguísticas particular de cada idioma. Enquanto na língua inglesa, muitas ferramentas estão disponíveis, em outros idiomas, ainda há bastante espaço para desenvolvimento.

O problema de escassez de recursos em outras línguas se estende para as ontologias de domínio que tem grande importância para a efetividade das abordagens OBIE.

4.4 Combinação de ontologias de domínio

(Wimalasuriya & Dou, 2010) discutem a combinação de ontologias de domínio como uma oportunidade para melhorar as abordagens de OBIE. Porém, percebemos que apenas algumas abordagens utilizam este recurso (Ali et al., 2017) (de Araujo et al., 2017).

Busca-se, portanto, explorar mais a interoperabilidade de ontologias para serem aplicadas em abordagens IE.

4.5 Desempenho da arquitetura OBIE

Uma das questões levantadas por (Wimalasuriya & Dou, 2010) é quanto ao desempenho dos sistemas de OBIE. Ainda há certa preocupação quanto aos recursos computacionais utilizados (tempo de processamento e memória) nas abordagens de OBIE. Para que os sistemas propostos sejam escaláveis, esses recursos precisam ser otimizados. O estudo (de Araujo et al., 2017) propôs o algoritmo ENI para reduzir a quantidade de memória computacional, bem como, o tempo necessário para processar os textos de linguagem natural. Takahashi, Mahmood, & Lakhani (2017) propõem uma arquitetura de *kernel cache* para sistemas de extração de informação multi-ontologia no Microsoft Windows para lidar com problemas de desempenho (Takahashi, Mahmood, & Lakhani, 2017).

A questão de desempenho se torna mais problemática com a utilização de abordagens de ML em OBIE. Os algoritmos de ML supervisionados necessitam lidar com grande quantidade de dados para aprender regras de extração.

4.6 Qualidade da ontologia

Há grande consenso sobre o fato de que eficácia e eficiência das abordagens OBIE dependem da cobertura e qualidade das ontologias (K. Liu & El-Gohary, 2017; Zhou & El-Gohary, 2017). Os estudos evidenciam que problemas na ontologia, como a ambiguidade, afetam a performance dos métodos de extração.

Os processos atuais ainda requerem muito esforço humano para gerar ou adaptar ontologias com qualidade. Há, inclusive, oportunidade para verificar quais são os pontos críticos na construção de ontologias de domínio que impactem as abordagens de OBIE.

5 Conclusão

Este trabalho mostrou um panorama sobre abordagens de extração de informação baseada em ontologia (OBIE). Para isso, relacionou abordagens de OBIE dos últimos anos através de um mapeamento sistemático. À partir de uma proposta de processo genérico (Figura 1), os trabalhos selecionados foram comparados com base no método de IE, baseados em regras e baseados em ML, bem como em características que impactam na forma como ontologias são utilizadas em OBIE (Tabela 1 e

Tabela 2). Vale ressaltar que mais eixos de comparação foram adicionados a essa tabela, em relação a principal referência de OBIE: (Wimalasuriya & Dou, 2010).

O mapeamento sistemático permitiu enxergar questões que ainda estão em aberto na área de OBIE. Levantamos, portanto, seis tópicos que oferecem desafios e oportunidades para sistemas de OBIE.

O primeiro diz respeito aos métodos de IE utilizados em OBIE. Há uma tendência que os métodos de ML sejam mais utilizados, pois devem diminuir o esforço humano para a geração de regras linguísticas, em contrapartida, os algoritmos de ML requerem mais recursos computacionais. A segunda questão é sobre a etapa de povoamento de ontologias. Apesar desta etapa ser uma premissa de sistemas OBIE, nem todas as abordagens fazem menção a desta. O terceiro tópico discute a falta de ferramentas e recursos para IE em idiomas diferentes do inglês, sendo este um limitante para o sucesso de OBIE em outras línguas. No quarto tópico, falamos sobre a combinação de ontologias para o aprimoramento de OBIE, sendo esta uma sugestão de (Wimalasuriya & Dou, 2010), mas que ainda é pouco explorada. O quinto tópico levanta a questão do desempenho em OBIE que se agrava com o uso de técnicas de ML. Finalmente, o sexto tópico diz respeito ao impacto da qualidade de ontologias em abordagens OBIE. Os autores defendem que problemas de ambiguidade na ontologia podem afetar gravemente na performance de OBIE.

Em trabalhos futuros, buscamos aprofundar técnicas de ML, identificar fatores críticos na construção da ontologia e explorar mais sobre essas questões que estão em aberto.

Referências Bibliográficas

Ali, F., Kwak, D., Khan, P., Ei-Sappagh, S. H. A., Islam, S. M. R., Park, D., & Kwak, K.-S. (2017). Merged Ontology and SVM-Based Information Extraction and Recommendation System for Social Robots. *IEEE Access*, *5*, 12364–12379. https://doi.org/10.1109/ACCESS.2017.2718038

Antoniou, G., & van Harmelen, F. (2004). *Handbook on Ontologies*. (S. Staab & R. Studer, Eds.), *Handbook on Ontologies*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24750-0

Bird, S. (2006). NLTK. In *Proceedings of the COLING/ACL on Interactive presentation sessions* - (pp. 69–72). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1225403.1225421

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkh061

Comerford, L., Frank, D., Gopalakrishnan, P., Gopinath, R., & Sedivy, J. (2001). The IBM Personal Speech Assistant. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* (Vol. 1, pp. 1–4). IEEE.

https://doi.org/10.1109/ICASSP.2001.940752

Community, A. O. D. (2011). *Apache OpenNLP Developer Documentation.pdf*. *OpenNLP*.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*. https://doi.org/10.1023/A:1014348124664

Cunningham, H., Maynard, D., & Tablan, V. (2000). *JAPE: a Java Annotation Patterns Engine*. *2000*.

de Araujo, D. A., Rigo, S. J., & Barbosa, J. L. V. (2017). Ontology-based information extraction for juridical events with case studies in Brazilian legal realm. *Artificial Intelligence and Law*, *25*(4), 379–396. https://doi.org/10.1007/s10506-017-9203-z

De Silva, N., Dou, D., & Huang, J. (2017). Discovering inconsistencies in PubMed abstracts through ontology-based information extraction. *ACM-BCB 2017 - Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 362–371. https://doi.org/10.1145/3107411.3107452

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*. https://doi.org/10.1017/S1351324904003523

Fudholi, D. H. D. H., Rahayu, W., & Pardede, E. (2016). Ontology-Based Information Extraction for Knowledge Enrichment and Validation. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)* (Vol. 2016–May, pp. 1116–1123). IEEE. https://doi.org/10.1109/AINA.2016.70

Gruber, T. (2007). Ontology. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (1st ed.). Springer-Verlag.

Jayawardhana, U. K., & Gorsevski, P. V. (2019). An ontology-based framework for extracting spatio-temporal influenza data using Twitter. *International Journal of Digital Earth*, *12*(1), 2–24. https://doi.org/10.1080/17538947.2017.1411535

Lima, R., Espinasse, B., & Freitas, F. (2017). OntoILPER: an ontology- and inductive logic programming-based system to extract entities and relations from text. *Knowledge and Information Systems*, *56*(1), 1–33. https://doi.org/10.1007/s10115-017-1108-3

Liu, K., & El-Gohary, N. (2016). Ontology-based Sequence Labelling for Automated Information Extraction for Supporting Bridge Data Analytics. *Procedia Engineering*, *145*, 504–510. https://doi.org/10.1016/j.proeng.2016.04.035

Liu, K., & El-Gohary, N. (2017). Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. *Automation in Construction*, *81*, 313–327. https://doi.org/10.1016/j.autcon.2017.02.003

Manine, A. P., Alphonse, E., & Bessières, P. (2008). Information Extraction as an Ontology Population Task and Its Application to Genic Interactions. In *2008 20th IEEE International Conference on Tools with Artificial Intelligence* (Vol. 2, pp. 74–81). https://doi.org/10.1109/ICTAI.2008.117

Mannai, M., Karâa, W. B. A. W. B. A., & Ghezala, H. H. B. H. H. Ben. Information extraction approaches: A survey, 625 Advances in Intelligent Systems and Computing § (2018). Springer, Singapore. https://doi.org/10.1007/978-981-10-5508-9_28

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Stroudsburg, PA, USA: Association for Computational Linguistics.

https://doi.org/10.3115/v1/P14-5010

Martinez, A., Yannuzzi, M., de Vergara, J. E. L., Serral-Gracia, R., & Ramirez, W. (2015). An Ontology-Based Information Extraction System for bridging the configuration gap in hybrid SDN environments. In *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)* (pp. 441–449). IEEE. https://doi.org/10.1109/INM.2015.7140321

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*. https://doi.org/10.1145/219717.219748

Pianta, E., Girardi, C., & Fondazione, R. Z. (2008). The TextPro tool suite. In *Proc. of the 6th Language Resources and Evaluation Conference (LREC 2008)*.

Rau, L. F., Jacobs, P. S., & Zernik, U. (1989). Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing and Management*, *25*(4), 419–428. https://doi.org/10.1016/0306-4573(89)90069-1

Rizvi, S. T. R., Mercier, D., Agne, S., Erkel, S., Dengel, A., & Ahmed, S. (2018). Ontology-based Information Extraction from Technical Documents. In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence* (Vol. 2, pp. 493–500). SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0006596604930500

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach, 3rd edition*. *Pearson*. https://doi.org/10.1017/S0269888900007724

Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012). Open Language Learning for Information Extraction. *Emnlp*.

Shawar, B. A., & Atwell, E. (2004). Accessing an Information System by Chatting. In *International Conference on Application of Natural Language to Information Systems* (pp. 407–412). https://doi.org/10.1007/978-3-540-27779-8_39

Takahashi, H., Mahmood, K., & Lakhani, U. (2017). Autonomous Decentralized Kernel Cache Architecture for Multi Ontology Based Information Extraction on Microsoft Windows. In *2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)* (pp. 15–22). IEEE. https://doi.org/10.1109/ISADS.2017.9

Viani, N., Larizza, C., Tibollo, V., Napolitano, C., Priori, S. G. S. G., Bellazzi, R., & Sacchi, L. (2018). Information extraction from Italian medical reports: An ontology-driven approach. *International Journal of Medical Informatics*, *111*(December 2017), 140–148. https://doi.org/10.1016/j.ijmedinf.2017.12.013

Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, *36*(3), 306–323. https://doi.org/10.1177/0165551509360123

Zhou, P., & El-Gohary, N. (2017). Ontology-based automated information extraction from building energy conservation codes. *Automation in Construction*, *74*, 103–117. https://doi.org/10.1016/j.autcon.2016.09.004

Tabela 1 - relação de abordagens recentes de OBIE – Métodos de IE, Construção e Componentes da Ontologia

Autor	IE baseado em regras	IE baseado em Aprendizado de Máquina	Construção de Ontologia	Povoamento de Ontologia	Componentes da ontologia
(Jayawardhana & Gorsevski, 2019)	Anotações léxico-sintáticas; anotações semânticas; similaridade Semântica	Não	Já existente	Não	Classes
(Viani et al., 2018)	Anotações léxico-sintáticas; anotações semânticas	Não	Construção com metodologia	Sim	Instâncias e valores de propriedades
(Rizvi et al., 2018)	Parser HTML; heurísticas	Não	Manualmente definido	Não	Classes
(De Silva et al., 2017)	Similaridade de palavras; regras linguísticas; POS; listas de gazetter; árvore *parse*	Não	Já existente	Não	Classes; Instâncias
(Lima et al., 2017)	Anotações léxico-sintáticas; anotações semânticas; classificação	Não-supervisionado	Construção pelo processo	Sim	Classes; relações; instâncias; valores de propriedades
(K. Liu & El-Gohary, 2017)	Anotações léxico-sintáticas; anotações semânticas; classificação; similaridade Semântica	Semi-supervisionado; *conditional-random fields*	Já existente	Não	Classes
(Ali et al., 2017)	Anotações léxico-sintáticas; anotações semânticas; Similaridade semântica	Supervisionado	Construção com metodologia	Não	Classes; instâncias; valores de propriedades
(Zhou & El-Gohary, 2017)	Classificação; anotações léxico-sintáticas; anotações semânticas; sequencial baseada em dependência e em cascata	Não	Construção com metodologia	Não	Classes
(de Araujo et al., 2017)	Anotações léxico-sintáticas; anotações semânticas	Semi-supervisionado	Construção com metodologia	Não	Classes; relações; instâncias; valores de propriedades
(Fudholi et al., 2016)	Anotações léxico-sintáticas; anotações semânticas	Não	Já existente	Sim	Classes; relações; instâncias; valores de propriedades

Tabela 2 - relação de abordagens recentes de OBIE – Fontes, idioma, ferramentas e avaliação

Autor	Fonte de dados	Idioma	Ferramentas/Frameworks/ Outras Tecnologias	Dicionários/*Thesaurus*/ *Gazetteers*	Validação da abordagem	Métricas de Avaliação
(Jayawardhana & Gorsevski, 2019)	Tweets do Twitter	Inglês	ArkTweet NLP; PostgreSQL; Postgis; GeoServer	DBpedia Spotlight	Experimento: Manual x Algoritmo	Precision, Recall, F-Measure
(Viani et al., 2018)	Documentos do domínio	Italiano + Inglês	TextPro (Pianta, Girardi, & Fondazione, 2008); UIMA Framework (Ferrucci & Lally, 2004); Protegé	UMLS Metathesaurus (Bodenreider, 2004), FederFarma	Estudo de caso: Validação em conjunto parcial/ *dataset* diferente	Precision
(Rizvi et al., 2018)	Documentos do domínio (PDF)	Inglês	Adobe Acrobat Pro	Não	Estudo de caso: Validação em conjunto parcial	Precision, Recall, F-Measure
(De Silva et al., 2017)	Artigos do PubMed[4]	Inglês	Stanford CoreNLP (Manning et al., 2014); OLLIE (Schmitz, Bart, Soderland, & Etzioni, 2012)	WordNet (Miller, 1995)+dicionário próprio	Não menciona	Não menciona
(Lima et al., 2017)	TREC Dataset[5]	Inglês	Stanford CoreNLP (Manning et al., 2014); OpenNLP (Community, 2011); Prolog	Não	*Fivefold cross-validation*; comparação com outros sistemas	Precision, Recall, F-1 Measure
(K. Liu & El-Gohary, 2017)	Documentos do domínio	Inglês	NLTK POS tagger (Bird, 2006)	Não	Experimento: Manual X Algoritmo; comparação com outros sistemas	Precision, Recall, F-1 Measure
(Ali et al., 2017)	Páginas da web em geral	Inglês	GATE (Cunningham, 2002); Google Search Engine; Protegé + plugins	Wordnet (Miller, 1995)	Experimento: Manual X Algoritmo	Precision, Recall, Accuracy
(Zhou & El-Gohary, 2017)	Documentos do domínio	Inglês	ANNIE (Cunningham, 2002); GATE (Cunningham, 2002); JAPE Transducer (Cunningham, Maynard, & Tablan, 2000)	ANNIE Gazetteer (Cunningham, 2002); OntoRoot Gazetteer (Cunningham, 2002)	Experimento: Manual x Algoritmo	Precision, Recall
(de Araujo et al., 2017)	Documentos do domínio	Port. BR	POWLA [Chiarcos]; OWL API; Java	Não	Estudo de caso: Manual x Algoritmo	Precision, Recall
(Fudholi et al., 2016)	Notícias do Google News	Inglês	GATE (Cunningham, 2002); JAPE Transducer (Cunningham et al., 2000); POS-tagger	WordNet (Miller, 1995); ontology gazetter (Cunningham, 2002)	Estudo de caso: comparação com outros sistemas	Precision

[4] https://www.ncbi.nlm.nih.gov/pubmed/

[5] http://cogcomp.cs.illinois.edu/Data/ER/conll04.corp.